DOCUMENT RESUME

ED 387 525 TM 023 985

AUTHOR Schmitt, Alicia P.; And Others
TITLE Evaluating Hypotheses about Differential Item Functioning.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-92-8
PUB DATE Jan 92
NOTE 70p.; Version of a paper presented at the Educational Testing Service/AFHRL Conference (Princeton, NJ, October 6, 1989).
PUB TYPE Guides - Non-Classroom Use (055) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Causal Models; Classification; Effect Size; *Estimation (Mathematics); *Hypothesis Testing; *Item Bias; Observation; *Regression (Statistics); Sampling; Test Construction; *Test Items
IDENTIFIERS Randomization; Standardization

ABSTRACT

Studies evaluating hypotheses about sources of differential item functioning (DIF) are classified into two categories: observational studies evaluating operational items and randomized DIF studies evaluating specially constructed items. For observational studies, advice is given for item classification, sample selection, the matching criterion, and the choice of DIF techniques, as well as how to summarize, synthesize, and translate DIF data into DIF hypotheses. In randomized DIF studies of specially constructed items, specific hypotheses, often generated from observational studies, are evaluated under rigorous conditions. Advice for these studies focuses on the importance of carefully constructed items to assess DIF hypotheses. In addition, randomized DIF studies are cast within a causal inference framework, which provides a justification for the use of standardization analyses or logistic regression analysis to estimate effect sizes. Two studies that have components spanning the observational and controlled domains are summarized for illustrative purposes. Standardization analyses are used for both studies. Special logistic regression analyses of an item from one of these studies are provided to illustrate a new approach in the assessment of DIF hypotheses using specially constructed items. (Contains 5 figures and 39 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 387 525

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

EVALUATING HYPOTHESES ABOUT DIFFERENTIAL ITEM FUNCTIONING

Alicia P. Schmitt
Paul W. Holland
Neil J. Dorans

Educational Testing Service
Princeton, New Jersey
January 1992

BEST COPY AVAILABLE

EVALUATING HYPOTHESES ABOUT DIFFERENTIAL ITEM FUNCTIONING[1,2]

Alicia P. Schmitt
Paul W. Holland
Neil J. Dorans

Educational Testing Service

[1]An earlier version of this report was presented at ETS/AFHRL Conference, *Differential Item Functioning: Theory and Practice*, Educational Testing Service, Princeton, NJ, October 6, 1989. This report will appear as Chapter 14 in P.W. Holland & H. Wainer (Eds.), *Differential Item Functioning*, which will be published by Lawrence Erlbaum Associates in 1992. Citations for this work should be to the chapter version.

[2]The opinions expressed in this report are those of one or all of the authors and should not be misconstrued to represent official policy of the Educational Testing Service. The authors are grateful to Howard Wainer and Charlie Lewis for their careful reviews and discussions of earlier versions of this report.

Copyright © 1992. Educational Testing Service. All rights reserved.

Abstract

Studies evaluating hypotheses about sources of differential item functioning (DIF) are classified into two categories: observational studies evaluating operational items and randomized DIF studies evaluating specially constructed items. For observational studies, advice is given for item classification, sample selection, the matching criterion, and the choice of DIF techniques, as well as how to summarize, synthesize and translate DIF data into DIF hypotheses. In randomized DIF studies of specially constructed items, specific hypotheses, often generated from observational studies, are evaluated under rigorous conditions. Advice for these studies focuses on the importance of carefully constructed items to assess DIF hypotheses. In addition, randomized DIF studies are cast within a causal inference framework, which provides a justification for the use of standardization analyses or logistic regression analysis to estimate effect sizes. Two studies that have components spanning the observational and controlled domains are summarized for illustrative purposes. Standardization analyses are used for both studies. Special logistic regression analyses of an item from one of these studies are provided to illustrate a new approach in the assessment of DIF hypotheses using specially constructed items.

EVALUATING HYPOTHESES ABOUT DIFFERENTIAL ITEM FUNCTIONING

Alicia P. Schmitt, Paul W. Holland, and Neil J. Dorans

Educational Testing Service

1. INTRODUCTION

Differential item functioning (DIF) research has had as one of its major goals the identification of the causes of DIF. Typically, DIF research has focused on determining characteristics of test items that are related differentially to subgroups of examinees and thus, which might explain or be a cause of DIF in an item. The premise has been, that, after identifying specific DIF-related factors, test development guidelines could be generated to prevent their future occurrence. With the elimination of these DIF factors, the items would not exhibit DIF and, thus, the total score would provide a better estimate of the true abilities of examinees from any subpopulation. The reality is that, to date, only a limited number of hypothesized DIF factors seem to hold consistently and that even these factors need to be better understood so that test construction guidelines can address them with the needed specificity.

There are several reasons why progress in the identification of DIF-related factors has been slow. First, the study of DIF is relatively new and so the initial emphasis was on the development of statistical methods to identify DIF. Dorans and Holland (in press) and Thissen, Steinberg, and Wainer (in press), provide good descriptions of the state-of-the-art statistical methods used to detect DIF.

Second, it requires a theory of differential item difficulty in a field in which theories of item difficulty are not well developed. Related to this is the fact that the reference and focal groups used to date--Blacks, Hispanics, Asians-Americans, and women, for example--are very heterogeneous and their differences are not easy to describe.

Third, the identification process is complex. Since more than one factor could be related to DIF in a given item, zeroing in on the specific cause of DIF for one item is not a simple process and confirming studies designed to test hypotheses about the causes of DIF are rare. We hope this paper helps to stimulate further empirical work in this area.

The purpose of this paper is to present and propose procedures for the systematic evaluation and corroboration of DIF-related factors or hypotheses. Descriptions of procedures to undertake observational DIF studies, to develop hypotheses, and to evaluate and construct items with the hypothesized factors are presented. Analytical comparison analyses are described and examples provided.

The systematic evaluation of DIF hypotheses involves a two-step process. The first step entails measuring DIF on regular operational items and using this information to generate hypotheses. The second step is a confirmatory evaluation of those hypotheses generated in step one. Thus, the main focus of the second step is the randomized DIF study in which specially constructed items are developed to test specific hypotheses and administered under conditions that permit appropriate statistical analyses to assess the efficacy of the hypotheses.

2. OBSERVATIONAL STUDIES: EVALUATING OPERATIONAL ITEMS

Hypotheses about factors related to DIF can be generated on the basis of theoretical or empirical considerations. Theoretical DIF hypotheses are founded on prior knowledge pertaining to cognitive processes that could be related to differential performance of test items. Although theoretical generation of DIF hypotheses is conceptually the first and most reasonable way to postulate logical reasons for DIF, it has not been very fruitful. Most test construction practices are carefully developed to avoid obvious factors that are known or suspected to be possible sources of discrimination toward any subpopulation of examinees. Processes such as the Test Sensitivity Review Process used at Educational Testing Service are used to evaluate developed items to ensure fairness to women and ethnic groups. This process is discussed by Ramsey (in press). Evaluation criteria for such sensitivity review procedures are designed so that items included in a test "...measure factors unrelated to such groups (minorities and women)" (Hunter & Slaughter, 1980, p.8). Therefore, logical or theoretical causes of DIF due to discrimination against women or ethnic minorities are supposed to be excluded from test instruments and thus can not be evaluated.

Empirical DIF hypotheses, generated after analyses of DIF data, may suggest that certain characteristics of items are differentially related to one or more subgroups of the population. Observational studies refer to investigations that make use of data and items constructed and administered under operational conditions. DIF analyses are conducted for all items in these tests to evaluate whether any item exhibits differential functioning by women or minority examinees. Performance of women on each item is compared to the performance of matched men (reference group for the female focal group) while the item performance of each minority

group (e.g., Asian-American, Black, Hispanic, and Native-American examinees) is contrasted to that of comparable White examinees (reference group for each minority focal group). Results of these DIF analyses can provide empirical information to generate DIF hypotheses.

2.1 DEVELOPMENT OF HYPOTHESES: EXTREME DIF ITEMS.

Evaluation of items with extreme DIF can provide insight into factors that might be related to DIF. Such a process involves a careful examination of the items with extreme DIF by a variety of experts. The speculation about or insight into possible causes of DIF for these items from test developers, researchers, focal group members, cognitive psychologists, and subject specialists can be used to generate hypotheses. Differential distractor information can engender additional insight into causes of DIF. Knowledge about which distractors differentially attract a specific subgroup may help to understand the respondents' cognitive processes. Differential distractor analyses are described by Dorans and Holland (in press) and Thissen, Steinberg, and Wainer (in press). Usually, analyses of more than one test form might be required in order to observe commonalities across items identified as having extreme differential performance. Some of the generated hypotheses might only consist of a general speculation or "story" about sources of DIF. It is important to consider any possible explanation. Since this stage is a generation-of-ideas phase, it can be considered almost a "brainstorming" process. Those possible explanations deemed most reasonable can then be developed into hypotheses to be tested.

2.2 EVALUATION OF HYPOTHESES THROUGH OBSERVATIONAL DATA

Once a number of possible hypotheses have been identified, the next step is to evaluate the efficacy of these hypotheses. Procedural steps to evaluate DIF hypotheses using observational data are delineated below.

Classification of Items

In order to evaluate the hypotheses, all items of a test form under study need to be classified with respect to the various hypothesized item factors or characteristics. A clear and precise definition of the factors to be studied needs to be provided. At least two experts or judges should classify each item according to each hypothesized factor. In cases where the two judges disagree, a third expert should be consulted. In this fashion, each item is identified as containing or not containing the factor or item characteristic under evaluation. Typically, a dichotomous classification is coded for each item factor. In those cases when a factor might consist of gradients or levels, a more continuous classification is appropriate. In addition, information about related variables might also be identified and coded. For example, the location of the item factor of interest (i.e., in the stem, key, or distractors) or the item type (e.g., antonyms, analogies, sentence completion, reading comprehension for verbal items) might provide information relevant to the relation of the factor to DIF. In fact, current research has shown that the greatest relationship between true cognates and DIF for Hispanic examinees is found when all components of an item have true cognates and the next greatest effect is found for those items with true cognates in the stem and/or key. On the SAT, these relationships were

found to be most notable for antonym and analogy item types (Schmitt, 1988; Schmitt, Curley, Bleistein, & Dorans, 1988; Schmitt & Dorans, 1990b).

Sampling Procedures

Groups

DIF factors can be postulated to be related to the differential item performance between two groups of examinees. In some instances, a postulated factor might not be specific to any one group. In such cases, more than one focal group might be of interest in a particular study. Typically, focal and reference groups have been determined on the basis of their gender and/or race or ethnic origin (i.e., females as focal group with males as reference group and Asian Americans, Blacks, Hispanics, or Native Americans as focal groups with Whites as reference group). Nevertheless, other characteristics (e.g., income level, educational background, or language knowledge) can serve to either further delimit ethnic or gender groups or to define other distinctive groups of interest. How focal and reference groups are determined and delimited depends both on the population characteristics of the examinees for whom the test is designed and intended as well as on hypothesized group characteristics. Cautious circumspection on the number of characteristics chosen to determine the groups under study is recommended. As the number of group-delimiting variables increases, the sample size of these groups is consequently restricted. Moreover, when several variables determine a group, findings about factors related to DIF are harder to interpret and their effect harder to ascribe to specific group variables.

Sample Size

All possible examinees on each focal and reference group should be used when doing DIF research. Because the comparison of comparable groups of examinees is an important component in the calculation of DIF statistics, differences on item performance of focal and reference groups are calculated at each ability level. Ability levels based on a predetermined criterion (e.g., total test score or another related ability measure) are used in the computation of DIF indices in a fashion analogous to how a blocking variable is used in a randomized block or in a split-plot design. For this reason, a reasonable number of examinees at each ability level is essential. The largest possible number of examinees in both the reference and focal groups should be used to render stable DIF estimates and to ensure sufficient power to detect DIF effects. The standard error of the DIF statistic should be examined to help interpret results when samples are small. Dorans and Holland (in press) and Donoghue, Holland, and Thayer (in press) discuss the standard error formulas and their accuracy.

DIF Analyses Procedures

Statistical Procedures

What statistical measure of DIF to use when conducting observational DIF studies is no longer the controversial decision it once was. The notable development and comparison of several DIF statistical methods during this decade have produced methods that, not only are reliable, but that generally have good agreement (Dorans, 1989; Dorans & Holland, in press; Dorans & Kulick, 1986; Holland, 1985; Holland & Thayer, 1988; Donoghue, Holland & Thayer, in press; Thissen, et al., in press; Scheuneman & Bleistein, 1989). Moreover, use of

more than one statistical method may be recommended. Currently, the operational assessment of DIF at Educational Testing Service uses the Mantel-Haenszel procedure to flag items for DIF and the closely related standardization procedure as a statistical tool to generate and assess content-based explanations for DIF. As mentioned previously, in addition to statistical indicants of DIF for the correct response, the development and evaluation of DIF hypotheses benefit from differential information on distractor selection, omitted responses, and speededness. Similarly, evaluation of empirical-option test regression curves and conditional differential response-rate plots for all these responses can indicate if any DIF effect is dependent on ability. Refer to Dorans and Holland (in press), and to Dorans, Schmitt, and Bleistein (1988) for descriptions of how to apply the standardization method to the computation of differential distractor, omit, and speededness functioning. Use of a log-linear model to examine DIF through the analysis of distractor choices by examinees who answered an item incorrectly is described by Green, Crone, and Folk (1989). Also see Thissen, et al (in press), for a discussion of differential alternative functioning (DAF).

Matching Criterion

The comparability of the focal and reference groups is achieved by matching these groups on the basis of a measure of test performance. Typically this measure is the total score on the test to be evaluated for DIF and is sometimes referred to as an internal matching criterion.

The major consideration in the selection of an appropriate DIF matching criterion is the degree of relationship between the construct of interest and the criterion. For DIF analyses, the construct of interest is what the test item is constructed to measure. If the total score matching

criterion is multidimensional, it will be measuring more than the construct of interest and may not be highly related to the item. Use of such a multidimensional total score criterion could compromise the comparability of the groups for a specific test item.

Another possible source of error in the estimation of a comparable total DIF matching criterion for the focal and reference groups is differential speededness (Dorans, Schmitt & Bleistein, 1988). Several studies are currently evaluating the effect of differential speededness and are considering this proposed speededness refinement (Schmitt, Dorans, Crone & Maneckshana, 1991).

Differential Response Style Factors

Different examinees approach the test taking experience differently and these different response style factors may have an impact on DIF assessment, particularly for items at the end of test sections. These response style factors are differential speededness and differential omission (Schmitt & Dorans, 1990). When an examinee does not respond to an item and does not respond to any subsequent items in a timed test section, all those items are referred to as "not reached". Differential speededness refers to the existence of differential response rates between comparable focal and reference group examinees to items appearing at the end of a test section. When an examinee does not respond to an item, but responds to subsequent items, that lack of response is referred to as an "omit". Differential omission refers to the occurrence of differential omit rates between comparable focal and reference group examinees. Adjustments for these differential response styles are important when evaluating DIF hypotheses because their occurrence can confound results.

Descriptive Statistics

After all items have been classified and DIF indices estimated, DIF summary statistics are computed for each level of the factor. The unit of analysis in this step is the item. Means, medians, minimum and maximum values, and standard deviations can help evaluate the impact of the postulated factor. Examination of all items classified as containing the factor under study allows for the identification of items that differ from the positive or negative pattern expected for the factor. Closer examination of such items can provide valuable information about possible exceptions to the expected effect. Although correlation analysis can render useful associative information, use of this descriptive statistic is limited. The dichotomous nature of most of the hypothesized factors and the limited number of naturally occurring items with such factors restrict the usefulness of statistical significance tests. Furthermore, the lack of controls particular to naturalistic studies also hampers the evaluation of DIF hypotheses. Nevertheless, naturalistic studies are a good first step, providing information valuable for the postulation of DIF hypotheses and data for their evaluation and refinement.

Confirmatory studies are a natural next step in the evaluation of DIF hypotheses. These studies require the construction of items with the postulated characteristics and use scientific methods to ensure that extraneous factors are controlled so that the factors of interest can be accurately evaluated.

3. EXAMPLES OF OBSERVATIONAL DIF STUDIES

Two observational studies which have evaluated DIF hypotheses previously postulated on the basis of DIF information will be described and findings reported.

3.1 LANGUAGE AND CULTURAL CHARACTERISTICS RELATED TO HISPANIC DIF

An in-depth analysis of the extreme DIF items for Hispanic examinees on one form of the Verbal Scholastic Aptitude Test (SAT-V) helped identify characteristics of the items that might explain the differential functioning by two Hispanic subgroups. Four hypotheses were generated about the differential item functioning of Hispanic examinees on verbal aptitude test items. These hypotheses were:

1. *True cognates, or words with a common root in English and Spanish, will tend to favor Hispanic examinees. Example: music (musica).*

2. *False cognates, or words whose meaning is not the same in both languages, will tend to impede the performance of Hispanic examinees. Example: enviable--which means "sendable" in Spanish.*

3. *Homographs, or words that are spelled alike but which have different meanings, will tend to impede the performance of Hispanic examinees. Example: bark.*

4. *Items with content of special interest to Hispanics will tend to favor their performance. Most special-interest items will tend to be reading*

comprehension item types. Example: reading passage about Mexican-American women.

For a detailed description of the study see Schmitt (1985, 1988).

Although some of the hypotheses were supported by the data (true cognates and special interest specific to the Hispanic subgroup), the low frequencies of false cognates and homographs, in the two test editions studied, precluded their evaluation. Construction of forms where the occurrence of the postulated factors could be controlled and, thus, evaluated was proposed as a follow-up to this investigation. The Schmitt et al., (1988) follow-up study remedied the limited naturally occurring item factors by developing items with these factors and administering them in non-operational SAT-V sections. Procedures and results of this confirmatory investigation are described and reported in section 5.1.

3.2 DIFFERENTIAL SPEEDEDNESS

In an effort to identify factors that might contribute to DIF, Schmitt and Bleistein (1987) conducted an investigation of DIF for Blacks on SAT analogy items. Possible factors were drawn from the literature on analogical reasoning and previous DIF research on Black examinees (Dorans, 1982; Echternacht (1972); Kulick, 1984; Rogers & Kulick, 1987; Rogers, Dorans & Schmitt, 1986; Scheuneman (1978); Scheuneman, 1981; and Stricker (1982)). Schmitt and Bleistein performed their research in two steps. Hypotheses about analogical DIF were developed after close examination of the three 85-item SAT-Verbal test forms studied by Rogers and Kulick (1987). Following these analyses, two additional test forms were studied to validate

the hypothesized factors. Standardization analyses were conducted in which the key, all distractors, not reached, and omits all served as dependent variables.

The major finding was that Black students do not complete SAT-Verbal sections at the same rate as White students with comparable SAT-Verbal scores. This differential speededness effect appeared to account for much of the negative DIF for Blacks on SAT Verbal analogy items. When examinees who did not reach the item were excluded from the calculation of the standardized response rate differences, only a few analogy items exhibited DIF.

Dorans, Schmitt and Bleistein (1988) set out to document the differential speededness phenomenon for Blacks, Asian-Americans and Hispanics on several SAT test editions, including some studied by Schmitt and Bleistein (1987). They found that differential speededness was most noticeable for Blacks and virtually nonexistent for Asian-Americans when compared with matched groups of Whites. A randomized DIF study to evaluate differential speededness under controlled conditions is described in section 5.3.

4. EVALUATING SPECIALLY CONSTRUCTED ITEMS

...we have not yet proved that antecedent to be the cause until we have reversed the process and produced the effect by means of that antecedent artificially, and if, when we do so, the effect follows, the induction is complete....(Mill, 1843, p.252)

In contrast to observational studies that evaluate operational data and can only draw associational inferences about DIF and item characteristics, well-designed randomized studies using specially constructed test items can be used to draw causal inferences about DIF and

postulated item factors. It is not until we can confirm an associated relation between item factors and DIF by verifying that the expected DIF is found on specially constructed items with these characteristics, that we can ascribe these factors to be a cause of DIF. The basic features of these randomized DIF studies is the use of control or comparison items and randomized exposure of examinees to these items. The purpose of this section is to describe procedures for constructing these items, for designing these systematic investigations, and for analyzing their results. Examples of two studies where these techniques have been applied are presented.

4.1 METHOD AND DESIGN

Variables

The variables used in randomized DIF studies may be described by the following terminology adapted from experimental design: response variables, treatment variables, and covariates. The dependent variable is the measure of the behavior predicted by a DIF hypothesis (e.g., choosing a predicted response). The treatment variable indicates the extent to which the item has the postulated DIF characteristics. Covariates are subject characteristics that are not affected by exposure to a particular treatment, i.e., measures of performance on related types of items or measures of education level and English language proficiency.

Instrument Development

Some of the treatments in a randomized DIF study consist of exposure to test items that have been specially constructed to disadvantage one or more groups of examinees. For this

reason, great care must be exercised, in the construction of these testing instruments, to conform to test sensitivity and human subject guidelines as well as to test specifications and sound test development practice. These constraints on the final testing instruments insure that operational scores are not affected by the study.

At least two levels of the treatment are needed in order to compare the effect of the DIF factor of interest. Thus, two versions of each item are developed. Ideally, these two items are identical in every respect but for the factor to be tested. The item that includes the postulated DIF characteristic is the "treated" level (t) of the treatment variable. The other item (the version that excludes the DIF factor) is the control level (c) of the treatment variable. The goal of the construction of these pairs of items is to make them as similar as possible except for the factor that is being tested. Parallelism of the two item versions is an important requirement that may allow us to infer that differences in the differential performance on these two items is caused by the difference between the items i.e., the postulated factor. Achieving parallelism of the item pairs is often difficult to do in practice because test questions are complex stimulus material and a change in one aspect of an item often entails other changes as well.

In order to issue parallelism, when constructing parallel items it is important to control the following item characteristics: difficulty, discrimination, location in the test, item type, content, and the location of the key and distractors. The number of items used to test each factor is also an important consideration because the unit of analysis is the item itself. Thus, it is desirable to construct several item pairs testing each factor.

The items in a pair are constructed to test a specific DIF hypothesis and are designed so that the treatment item (t) is more likely to elicit a particular type of response than is the control

item (c) in the pair, especially for members of the focal group of interest. Schmitt et al. (1988) give the following example of a pair of specially constructed antonym items used in their randomized DIF study discussed further in section 5.1.

Item t	Item c
PALLID:	ASHEN:
(A) moist	(A) moist
(B) massive	(B) massive
(C)* vividly colored	(C)* vividly colored
(D) sweet smelling	(D) sweet smelling
(E) young and innocent	(E) young and innocent

These two items are identical except for their stem, i.e. PALLID or ASHEN, which are synonyms of roughly equal frequency in English. The factor being varied in this item pair is the existence of a Spanish cognate for the stem word. In this case, palido is a common word in Spanish while the cognate, pallid, is a less common word in English. The DIF hypothesis for this item pair is that the existence of a common Spanish cognate for a relatively more rare English word that plays an essential role in the item (in this case the stem word) will help Spanish speaking examinees select the correct answer and will not help non-Spanish speaking examinees.

Samples

The relevant reference and focal groups are determined by the DIF factors that are postulated. In a randomized DIF study, the reference and focal groups are then subdivided at

random into subgroups of examinees who are exposed to either the *t* or *c* version of each item. Because of this subdivision, it is important to have large samples of each of the target groups.

Controls

Randomized DIF studies are by definition controlled studies. The use of control or comparison items allows us to infer that differences in DIF between the item pairs is caused by differences between the items (the postulated factor) as long as other causative variables are not contaminating the results. There are three major types of extraneous variables that can contaminate results if not controlled: examinee related differences, lack of parallelism of the item pairs, and differences in the testing conditions of examinees taking each of the item pairs. The control of these extraneous variables needs to be carefully considered. The types of controls that can be used for this purpose are: randomization, constancy, balance, and counterbalance.

Randomization

If the examinees who are exposed to the *t* version of an item pair differ in important ways from those exposed to the *c* version, confounding is said to occur. Confounding makes it difficult to separate the effect of responding to the *t* versus the *c* version of the item pair from the characteristics of the examinees in these two groups. Randomization tends to equalize the distribution of examinee characteristics in the *t* and *c* groups. It may be achieved by spiralling subforms together each containing only one number of each pair of special items. We discuss the effect of randomization more formally in section 5.2.

Constancy

Some factors, such as the position of the key in a multiple choice question, can affect examinee responses and should be the same for items *t* and *c* in a given pair. This is an example of constancy. Other factors that might be controlled using this technique are item position and response options. The PALLID/ASHEN item pair is an example with a great deal of constancy across the pair of items. A special case of constancy arises when a factor is eliminated in the sense that it is prevented by design from occurring. An example of a factor that can be eliminated in a randomized DIF study is differential speededness. It is removed by placing the specially constructed items at the beginning of the test section.

Balance

Balance is used in two distinct ways. On the one hand, it can refer to equalizing the distribution of important examinee characteristics across the *t* and *c* versions of an item pair. Randomization will approximately balance the distribution of covariates in a large study, but in a small study the researcher may need to achieve balance in a more active way (i.e., blocking). On the other hand, balance can refer to the entire set of stimulus material that an examinee is exposed to. Subforms are usually balanced with respect to content and item type so that they will not appear unusual to the examinees and thereby will not elicit unrepresentative responses.

Counterbalance

Counterbalance refers to the stimulus material presented to the examinees. A factor that is appropriate to counterbalance across the subforms used in the study is the total number of occurrences of *t* and *c* items in each subform. When this is counterbalanced, all the subforms will have the same number of *t* and *c* items even though they cannot contain the same *t* and *c* items. This will tend to reduce the overall effect of each subform on differences in subgroup performance.

Other Considerations

The form of control used has an effect on the generality of the inferences made from the study. For example, if only one level of item difficulty is used in the evaluation of an hypothesis (i.e., constancy) then any resulting effect of the hypothesized DIF factor under study may be restricted to items with the tested level of difficulty. It is, therefore, important to select the method of control (i.e., balancing, etc.) based on the level of generality that is desired.

Another way to deal with extraneous variables is to control them in the design of the study. In such cases the DIF factor will be one independent variable while another variable, such as item difficulty, will be a second independent variable. In this example, we develop item pairs that have similar difficulty within a pair, but varying difficulty across pairs. When more than one independent variable is being studied at one time, evaluation of their interactive effect is a part of the study. If an interaction is found then analyses should proceed to see how the effect of the DIF factor varies across the levels of the interacting variables. Because the

outcome measures in DIF studies are usually response probabilities, the scale in which these are measured, P or logit, may affect the study of interactions.

A major constraint on using several independent variables in designing the item pairs is that the number of items has to be increased accordingly and the study made more complex. Several items need to be constructed to evaluate each possible combined effect. For example, if the DIF factor under study consists of two levels (DIF factor present or absent) and the other independent variable consists of three levels (e.g., item difficulty: hard, medium, and easy) then the total number of subgroups of items testing all possible combinations is six. If there are then at least two examples of each item there are at least 12 items to study a single DIF factor. Because of practical testing constraints, it may be necessary to limit the number of independent variables to be studied at a time in a randomized DIF study.

4.2 A CAUSAL INFERENCE PERSPECTIVE ON RANDOMIZED DIF STUDIES

This section adapts the formal model of Rubin (1974) and Holland (1986) to the analysis of randomized DIF studies.

Dependent Measures

In a randomized DIF study the basic dependent measure is the response an examinee gives to the specially constructed test items. Assuming that we are considering multiple choice tests, the responses of examinees are limited to choosing one of the response options or omitting the item. It is also possible that an examinee might not attempt to respond to some items but our analysis will condition on responding to the special items. Depending on exactly what DIF

hypothesis is being tested, the particular response of interest will vary. For many hypotheses the behavior of interest is choosing the correct answer. For others it might be choosing or not choosing a particular distractor and for others it will be the decision to omit the item. We will not make an assumption on this and will let the outcome variable, Y, be dichotomous with

$$Y = \begin{cases} 1 & \text{if the examinee makes the predicted response relevant to the DIF hypothesis,} \\ 0 & \text{otherwise} \end{cases}$$

(In all of our examples, however, we use $Y = 1$ to denote choosing the correct answer to the special items.)

There are two potential responses that could be observed for an examinee, Y_t, or Y_c, where

Y_t = the value of Y that will be observed if the examinee is asked to respond to item t of the pair,

Y_c = the value of Y that will be observed if the examinee is asked to respond to item c of the pair.

The difference, $Y_t - Y_c$, is the "causal effect" for a given examinee of being asked to respond to item t rather than to item c in the pair. Let S denote the member of the pair of special items to which the examinee is asked to respond, i.e. $S = t$ or $S = c$. Then Y_S is the actual response that the examinee gives in the study. The notation Y_S means the following:

$$Y_S = \begin{cases} Y_t \text{ if } S = t \\ Y_c \text{ if } S = c \ . \end{cases}$$

The problem of causal inference in a randomized DIF study is to say as much as we can about the unobservable causal effect, $Y_t - Y_c$, for each examinee from the observable data. For example, if $Y_t - Y_c = 0$ then the examinee would make the same response regardless of the version of the item to which he or she is exposed. When $Y_t - Y_c = 1$ then the examinee would make the predicted response to t but not to c, etc.

The Data

So far we have mentioned two pieces of information that are available from each examinee with respect to a given pair of items, the observed response Y_S and the member of the pair of special items to which the examinee responded, S. In addition there is other important information. First of all, the examinee may belong to the focal or the reference group of interest, or possibly to neither one. Denote group membership by $G = r$ or f (reference or focal). In addition there may be other test scores available for the examinee. For example, the special items may be part of a larger test. Let X denote an additional score obtained from part or all of this larger test. We must distinguish two important cases. If it is possible to assume that the score X is unaffected by whether or not the examinee was asked to respond to item t or to item c of the item pair of interest then X is called a covariate score. For example, if the items that are part of the X-score are all asked prior to the examinees being asked to respond to the special items then it is usually plausible to assume that X is a covariate score. On the other hand, if the special item is included in the X-score, then X is not a covariate score. We will use covariate scores to group examinees.

In summary, the data observed for a given examinee can be expressed as

Y_s, S, G, and X.

The Average Causal Effect

The individual level causal effect, $Y_t - Y_c$, is not directly observable for a single examinee because we only observe Y_t or Y_c (but not both) on each examinee. An average causal effect (ACE) is found by averaging the individual level causal effects over various groups of examinees. For example we might consider

$$E(Y_t - Y_c), \tag{1}$$

the average over everyone in the study, or

$$E(Y_t - Y_c \mid G = f), \tag{2}$$

the average over everyone in the focal group, or

$$E(Y_t - Y_c \mid G = r), \tag{3}$$

the average over everyone in the reference group. Finally we might consider

$$E(Y_t - Y_c \mid G = g, X = x), \tag{4}$$

the average over everyone in group g with covariate score x. We will show later that average causal effects can be estimated by the data obtained in a randomized DIF study, even though the examinee-level causal effects can not be estimated.

Let us consider the ACEs in (1) - (4) further. Because Y_t and Y_c are dichotomous, the expectations in (1) - (4) may be expressed in terms of probabilities, i.e.

$$E(Y_t - Y_c) = P(Y_t = 1) - P(Y_c = 1), \quad (5)$$

$$E(Y_t - Y_c \mid G = g) = P(Y_t = 1 \mid G = g) - P(Y_c = 1 \mid G = g), \quad (6)$$

$$E(Y_t - Y_c \mid G = g, X = x) = P(Y_t = 1 \mid G = g, X = x) - P(Y_c = 1 \mid G = g, X = x). \quad (7)$$

The ACE, $E(Y_t - Y_c)$, averages over all examinees and as such represents the "main effect" of item t relative to item c over all examinees. While this main effect is important, it is not the primary parameter of interest in a randomized DIF study. When Y represents choosing the correct option in a multiple choice test, the main effect (5) is simply the difference in the percent correct for items t and c over the examinees in the study. As we shall see, it is desirable to construct t and c so that the main effect (5) is small.

In general, the idea behind a randomized DIF study is that item t will elicit a bigger change in the probability of the predicted response relative to item c for members of the focal groups then it does for members of the reference group. This leads us to examinee the ACE-difference or interaction parameter defined by

$$T = E(Y_t - Y_c \mid G = f) - E(Y_t - Y_c \mid G = r), \quad (8)$$

$$= [P(Y_t = 1 \mid G = f) - P(Y_t = 1 \mid G = r)] - [P(Y_c = 1 \mid G = f) - P(Y_c = 1 \mid G = r)]. \quad (9)$$

It is useful to remember the two ways of writing T in (8) and (9). In (8) T is expressed as the difference between the ACE in the focal group and ACE in the reference group. In (9) T is expressed as the difference between the t and c items in their respective differences in the probability that $Y = 1$ between the focal and reference groups. When $Y = 1$ indicates a correct answer, the difference in the probability that $Y = 1$ between the focal and reference groups is called the impact of the item (Holland, 1985). Thus, in this case T may be viewed as the difference in the impact of item t and item c.

When T in (8) is positive it means that the change in the probability of the predicted responses caused by t (relative to c) is larger for the focal group than it is for the reference group (i.e., the ACE for f is larger than the ACE for r). Typically, this is the type of prediction made in a DIF hypothesis.

One problem with a parameter like T is that the probability of the predicted behavior measured by Y_t or Y_c will often differ between the reference and focal group, that is $P(Y_j = 1 \mid G = r)$ and $P(Y_j = 1 \mid G = f)$ will not be the same. It may also differ between item t and c, i.e. if there is a "main effect" of items in the pair. When these differences are large, the interpretation of the magnitude of T is complicated by the boundedness of the probability scale (i.e., the fact that Y is a 0/1 variable). Consider these four examples in which Y denotes selection of the correct response for a pair of items, (t, c).

Example A: $P(Y_t = 1 \mid G = r) = P(Y_c = 1 \mid G = r) = P(Y_t = 1 \mid G = f) = .5$,

and $P(Y_c = 1 \mid G = f) = .4$

then

$$T = (.5 - .4) - (.5 - .5) = .1.$$

The ACE in the reference group is 0 while in the focal group it is .1, so that T is .1. In this case items c and t are equally difficult for the reference group and t is equal in difficulty for the reference and focal groups. Furthermore, c is more difficult than t for the focal group. This is an ideal type of example in which some characteristic of item c causes it to be harder for members of the focal group and when this is altered to item t the item is equally difficult for both the reference and focal groups.

Example B: $P(Y_t = 1 \mid G = r) = .55$, $P(Y_c = 1 \mid G = r) = .45$,

$P(Y_t = 1 \mid G = f) = .50$, $P(Y_c = 1 \mid G = f) = .30$.

then the ACE for r is .1 and the ACE for f is .2, so that $T = .2 - .1 = .1$, again.

This is a more realistic example than example A because there is both a group difference and an item difference. Still it is evident in this example that the change from item c to item t has a bigger average causal effect on members of the focal group than it has for members of the reference group.

Example C: $P(Y_t = 1 \mid G = r) = .95$, $P(Y_c = 1 \mid G = r) = .85$

$P(Y_t = 1 \mid G = f) = .70$, $P(Y_c = 1 \mid G = f) = .50$

then

$T = (.70 - .50) - (.95 - .85) = .20 - .10 = .1$, once more.

The value of T is the same as in examples A and B but the context is quite different. Both c and t are much easier for the reference group than they are for the focal group and for both groups item t is somewhat easier than item c. The ACE for f is .70 - .50 = .20 but the ACE for r is only .95 - .85 = .10. However, the boundedness of the probability scale makes it impossible for $P(Y_t = 1 \mid G = r)$ to exceed $P(Y_c = 1 \mid G = r)$ by .20 when the latter probability is .85, as in this example. Does $T = .1$ mean that the change from c to t had a bigger effect for members of f then for members of r or was c already so easy for members of r that the change to t could not improve their performance as much as it did for members of f? This ambiguity stems from the large difference in performance on c and t between the reference and focal group. The use of covariate scores is aimed at removing some of this confusion--as we discuss below.

Example D: $P(Y_t = 1 \mid G = r) = .95$, $P(Y_c = 1 \mid G = r) = .60$,

and

$P(Y_t = 1 \mid G = f) = .85$, $P(Y_c = 1 \mid G = f) = .40$.

In this case

$$T = (.85 - .40) - (.95 - .60) = .1$$

as in the other examples. This example is like Example C except that the roles of the groups and the items have been reversed. In this example, there is a large main effect of items--item t is much easier for both groups than is item c. The consequence of this large main effect is that it confuses the interpretation of T. The ACE for f is .85 - .40 = .45 while the ACE for r is .95 - .60 = .35, however, starting with $P(Y_c = 1 | G = r) = .60$ it is impossible for the ACE for r to exceed .40. Again the boundedness of the probability scale is a source of confusion in the interpretation of T.

The Use of Covariate Scores

Examples C and D show that the boundedness of the probability scale can confuse the interpretation of the parameter T when there are large differences between the reference and focal groups in their probabilities of producing the predicted response for items t and or when there is a large main effect of items. The introduction of a covariate score can help alleviate this problem when there are large group differences. Large main effects of items are generally a sign of a poorly designed item pair for a DIF study.

Suppose X is a covariate score in the sense described earlier, i.e., X is measured on each examinee in the study and is not affected by exposure of the examinee to items t or c. Suppose further that examinees with the same X-score have similar probabilities of making the predicted

response to items t and c regardless of whether they are in the reference or focal group. That is, suppose that $P(Y_t = 1 | G = r, X = x)$ and $P(Y_t = 1 | G = f, X = x)$ are similar in value and that $P(Y_c = 1 | G = r, X = x)$ and $P(Y_c = 1 | G = f, X = x)$ are also similar in value. This latter assumption is what we mean by a useful covariate score. If the predicted behavior is choosing the correct response to items t and c then candidates for useful covariate scores are number right or formula scores based on sets of items that measure the same ability that is measured by items t and c. When the predicted behavior is choosing a particular distractor or omitting the item, number right or formula scores on other items may not produce a sufficiently useful covariate score and it may be necessary to augment test score with other variables, or to define scores based on special choices of distractors.

When X is a covariate score we can examine a third parameter based on the ACEs in (7). Define $T(x)$ by

$$T(x) = \mathrm{E}(Y_t - Y_c \mid G = f, X = x) - \mathrm{E}(Y_t - Y_c \mid G = r, X = x), \tag{10}$$

$$= [P(Y_t = 1 \mid G = f, X = x) - P(Y_c = 1 \mid G = f, X = x)]$$

$$- [P(Y_t = 1 \mid G = r, X = x) - P(Y_c = 1 \mid G = r, X = x)]. \tag{11}$$

The causal parameter, $T(x)$, is an interaction like T but is conditional on each X-score. When X is a useful covariate score and the main effect (1) is small the four probabilities in (11) will be similar and the boundedness of the probability scale will not confuse the interpretation of $T(x)$ to the degree that it can for T.

Even though $T(x)$ can help clarify the results of a comparison of responses to t and c for members of the reference and focal group, it does introduce the added complexity of a whole set of parameter values, one for each value of X, rather than just a single value. When X is a univariate score this plethora of parameters can be handled by a graph of $T(x)$ versus x. When X is a multivariate set of covariate scores this solution is not as helpful.

One way around this plethora of parameters is to average $T(x)$ over some distribution of X-values, $w(x)$, where $w(x) \geq 0$, $\sum_x w(x) = 1$. This results in a new parameter T_w defined by

$$T_w = \Sigma\, T(x)\, w(x) \tag{12}$$

$$= \sum_x [P(Y_t = 1 \mid G = f, X = x) - P(Y_t = 1 \mid G = r, X = x)]\, w(x)$$

$$- \sum_x [P(Y_c = 1 \mid G = f, X = x) - P(Y_c = 1 \mid G = r, X = x)]\, w(x). \tag{13}$$

The choice of $w(x)$ matters, and is somewhat arbitrary. In the standardization DIF procedure (Dorans & Holland, in press), the distributions of X in the focal group is often used as weights, i.e.,

$$w(x) = P(X = x \mid G = f).$$

This leads to the parameter that we denote by T_f given by

$$T_f = \sum_x T(x)\, P(X = x \mid G = f) \tag{14}$$

$$= \sum_x [P(Y_t = 1 \mid G = f, X = x) - P(Y_t = 1 \mid G = r, X = x)]\, P(X = x \mid G = f),$$

$$- \sum_x [P(Y_c = 1 \mid G = f, X = x) - P(Y_c = 1 \mid G = r, X = x)]\, P(X = x \mid G = f). \tag{15}$$

If we let

$$\Delta_t = \sum_x [P(Y_t = 1 \mid G = f, X = x) - P(Y_t = 1 \mid G = r, X = x)]\, P(X = x \mid G = f) \quad (16)$$

and

$$\Delta_c = \sum_x [P(Y_c = 1 \mid G = f, X = x) - P(Y_c = 1 \mid G = r, X = x)]\, P(X = x \mid G = f) \quad (17)$$

then

$$T_f = \Delta_t - \Delta_c. \quad (18)$$

In the case where X is a number right or formula score and the predicted behavior is selecting the correct response for items t and c, Δ_t and Δ_c are the parameters estimated by the standardization DIF procedure. Hence, T_f may be interpreted as the difference between standardization DIF parameters for items t and c.

At this point, it is worth stopping for a moment and asking why do we pay so much attention to the ACE parameters given in (1) - (4). After all, in computing a DIF measure for an item we compare the performance of matched focal and reference group members on the studied item and this is not an ACE parameter. To make the comparison sharper, in computing a DIF measure for an item using the standardization methodology the basic parameters are the differences

$$P(Y_j=1 \mid G = f, X^* = x) - P(Y_j = 1 \mid G = r, X^* = x) \quad (19)$$

for a fixed item $j = t$ or c and a score X^* that includes the score on the studied item 1[1]. In contrast, the corresponding ACE is

$$P(Y_t = 1 | G = g, X = x) - P(Y_c = 1 | G = g, X = x) \tag{20}$$

where g is a particular group, reference or focal, and X is covariate score that does not include the studied items.

The motivation for our emphasis on the ACE parameters is a causal model that underlies the observations. Consider the joint distribution of the two variables (Y_t, Y_c) over the set of examinees for which $G = g$ and $X = x$. Let this (conditional) joint distribution be denoted by

$$p_{uvgx} = P(Y_t = u, Y_c = v | G = g, X = x). \tag{21}$$

Thus, for example, p_{10gx} is the probability that a focal group member with covariate score $X = x$ will give the predicted response if responding to item t but will not give it if responding to item c. In this sense, then p_{10gx} is the probability that item t causes the predicted response for focal group examinees with covariate score $X = x$. The values of p_{uvgx} are "causal parameters" in this special, but clear-cut, sense. Notice that

$$P(Y_t = 1 | G = g, X = x) = p_{11gx} + p_{10gx} \tag{22}$$

[1]See Holland and Thayer (1988) and Donoghue, Holland and Thayer (in press) for a discussion of why inclusion of the studied item in the matching variable is important for both the Mantel-Haenszel and the standardization procedures.

Hence, the ACE parameter given in (4) can be expressed in terms of the causal parameters in (21) in the following way:

$$\begin{aligned} &E(Y_t - Y_c | G = g, X = x) \\ &= (p_{11gx} + p_{10gx}) - (p_{11gx} + p_{01gx}) \\ &= p_{10gx} - p_{01gx}. \end{aligned} \tag{23}$$

Finally, this gives us an important formula that relates the conditional-on-X ACE-difference, $T(x)$, to the causal parameters, i.e.,

$$T(x) = p_{10fx} - p_{01fx} - (p_{10rx} - p_{01rx}) \tag{24}$$

Equation (24) can be used to justify our emphasis on the ACE parameter in the following way. Suppose item c is just as likely to cause members of f to make the predicted response as it is to cause members of r to do this for examinees with $X = x$. This means that

$$p_{01fx} = p_{01rx} \tag{25}$$

It follows that if (25) holds then

$$T(x) = p_{10fx} - p_{10rx}, \tag{26}$$

so that in this case $T(x)$ is the excess of the probability that t causes the predicted response in the focal group over this probability in the reference group. Assumptions about the causal

parameters, p_{mgx}, are generally untestable, but, depending on the degree of control exercised in the design of the (t, c)-item pair, some assumptions can be made plausible and then give a direct causal interpretation to $T(x)$. We emphasize that (25) is not the only type of assumption that can arise in a randomized DIF study.

5. EXAMPLES OF SPECIAL CONFIRMATORY STUDIES

Randomized DIF that grew out of the two examples of observational studies discussed in Section 3 are described in this section. These studies either constructed items with the postulated factors or varied the location of the items. Other examples of DIF research evaluating effects of specially constructed items are: Bleistein, Schmitt, and Curley (1990) and Scheuneman (1984, 1987).

5.1 SYSTEMATIC EVALUATION OF HISPANIC DIF FACTORS

The purpose of the Schmitt et al. (1988) investigation was to provide a follow-up to the Schmitt (1985, 1988) studies through analysis of specially constructed SAT-V items in which the occurrence of postulated factors (true cognates, false cognates, homographs, and special interest) was rigorously controlled and manipulated. Two parallel 40-item non-operational sections were constructed so that each item in one form is a revised but very similar version of the same position item in the other form. The standardization method was used to compare the performance of the White reference group and each Hispanic focal group for each item in each of the two special forms. An external matching criterion was used, the 85-item SAT-V

operational examination taken in the same booklet as the specially constructed section under study. Hence, the studied items were not included in the matching criterion, which is the appropriate course of action for a randomized DIF study. Refer to section 4.2 for an explanation of why a studied item is not included as part of the matching criterion or covariate. Estimations of DIF were corrected for speededness by including only those examinees who reached the item in its calculation. In addition to calculating DIF values for the key, differences in the standardized proportion of responses for each distractor were computed and evaluated in order to further understand the effects of the hypothesized factors on Hispanic DIF. Empirical-option response curves and conditional differential response-rate plots were also evaluated for each item comparison.

Comparison of the DIF value obtained for one item version versus the DIF value obtained for the other item version indicated whether or not the postulated factor effect was supported or not. The most convincing support was found for the hypothesis that the true cognates facilitate the performance of Hispanic examinees. Striking effects were found for two antonym item pairs where the true cognates produced positive DIF values that exceeded 10% for nearly all Hispanic subgroups while the DIF value for the alternate neutral item indicated that the Hispanics groups performed slightly worse than the reference White group. The PALLID/ASHEN item pair (#7) presented in section 4.1 was one of these two antonym item pairs. Figure 1 presents differences in standardization DIF values between the item pairs testing the true cognate factors for the total Hispanic group. Confidence bands are drawn on this figure to indicate that differences greater than 3% between the DIF values of the item pairs are statistically significant. Although only the two antonym item pairs had differences (.17 and .15 for all Hispanics) that fell outside the

confidence band for all the Hispanic subgroups, some of the other item pairs had differences in the postulated direction.

Insert Figure 1 about here

Comparison of the true cognates with differences in the postulated direction versus those with no apparent DIF effect indicate that the true cognates that consistently made the items differentially easier for Hispanics were words with a higher frequency of usage in the Spanish language. Because of these results, the true cognate hypothesis was revised to restrict the positive effect of true cognates to true cognates with a higher usage in the Spanish language than their usage in the English language (Schmitt & Dorans, in press). Since mixed or marginal results were found for the other hypothesized factors the authors counseled:

> *More research is needed before prescriptive or proscriptive rules can be devised to guide item writers. The true cognate items demonstrate clearly, however, that DIF can be manipulated, at least some of the time. (Schmitt et al., 1988, p. 20)*

5.2 USING LOGISTIC REGRESSION TO ESTIMATE EFFECT SIZES

Section 4.2 discussed the parameters of interest in a randomized DIF study at the population level but did not discuss the details of how to estimate them. We now consider the problem of estimation. There are two parts to this discussion. The first concerns how random assignment of the special items to examinees allows the basic probabilities in (5) -(7) to be

estimated from the data collected in a randomized DIF study. The second concerns how to use modern discrete data models to estimate the causal parameters of interest. We discuss each in turn. We use the data from the randomized DIF study for Hispanics, described in Section 5.1, to illustrate how the procedure is used and to compare its estimates of effect size to those produced by standardization.

Randomization and the Causal Parameters

To reiterate, the various ACEs defined in (1) - (4) and the ACE-difference or interaction parameters, T and $T(x)$, defined in (8) and (10) are based on these probabilities:

$$P(Y_j = 1),\ P(Y_j = 1 \mid G = g) \text{ and } P(Y_j = 1 \mid G = g, X = x) \qquad (27)$$

for $j = c, t$ and $g = f, r$.

However, the data that is obtained in a randomized DIF study is Y_S, S, G and X on each examinee. Hence the parameters that can be directly estimated in a randomized DIF study are not those in (27), but are, instead,

$$P(Y_S = 1 \mid S = j),\ P(Y_S = 1 \mid S = j, G = g) \text{ and } P(Y_S = 1 \mid S = j, G = g, X = x), \qquad (28)$$

which can also be expressed as

$$P(Y_j = 1 \mid S = j),\ P(Y_j = 1 \mid S = j, G = g) \text{ and } P(Y_j = 1 \mid S = j, G = g, X = x). \qquad (29)$$

(Note that in (28) and (29) we have made use of the fact that X is a covariate score -- otherwise it would be subscripted by j.)

The role of random assignment of examinees to item t or c is that it makes the variable S statistically independent of Y_t, Y_c, G and X. Hence, randomization results in the probabilities in (29) being respectively equal to those in (27) that underlie the ACEs and ACE-differences of interest to us. Thus, we may use estimates of the probabilities in (28) as the basis of our inferences of the causal parameter T, $T(x)$, and T_w. If random assignment fails for some reason then this is not true. There are a variety of ways that random assignment can fail to be executed in any randomized study. An important class of such failure is "differential dropout" between the units assigned to each condition. In randomized DIF studies "drop-out" means that the examinee does not attempt to answer the special test items. Differential drop-out might occur between examinees assigned to item t and to item c if the location of these items in the overall test form is very different--i.e. t is the first item in its test form but c is the last item of its test form.

Estimating the Main Effect Parameter

Useful estimation strategies always depend on the type and extent of the available data. We will describe an approach, based on logistic regression, that can be used in a variety of situations. The main effect parameter

$$E(Y_t - Y_c) = P(Y_t = 1) - P(Y_c = 1)$$

can also be expressed, by the argument given above, as the treatment-control difference,

$$P(Y_t = 1 \mid S = t) - P(Y_c = 1 \mid S = c), \tag{30}$$

in a randomized DIF study. Let $\hat{p}_t$ denote the proportion of examinees who made the predicted response among those asked to respond to item t and let $\hat{p}_c$ be similarly defined for item c. Then the difference, $\hat{p}_t - \hat{p}_c$, estimates the difference in (30). For example, consider the PALLID/ASHEN item discussed earlier. A sample of 42,033 White or Hispanic examinees answered the PALLID item (t) and 45,960 White or Hispanic examinees answered the ASHEN item (c). The proportions answering the two items correctly are, respectively, .51 and .50. The estimate of the main effect of items is the difference, .01. Thus, we see that, in fact, the two items are nearly of equal difficulty, over the subpopulation consisting of proportional representations of self-identified White and Hispanic examinees. In this sample there were 84,852 White examinees and 3,141 Hispanic examinees.

It is useful to set up our notation for logistic regression now so that we can show its relationship to the main-effect parameter (30). Let S^* and G^* be indicator variables defined by

$$S^* = \begin{cases} 1 & \text{if } S = t, \\ 0 & \text{if } S = c, \end{cases} \quad , \quad G^* = \begin{cases} 1 & \text{if } G = f, \\ 0 & \text{if } G = r. \end{cases} \tag{31}$$

We set up a logistic regression model of the following fairly general form:

$$\text{logit } [P(Y_S = 1 \mid S, G, X)]$$

$$= \alpha_0 + \sum_{k=1}^{a} \alpha_k X^k + \beta_0 S^* + \gamma_0 G^* + \lambda_0 S^* G^*$$
$$+ \sum_{k=1}^{a} \beta_k S^* X^k + \sum_{k=1}^{a} \gamma_k G^* X^k + \sum_{k=1}^{a} \lambda_k S^* G^* X^k, \quad (32)$$

where

$$\text{logit}(p) = \log\left[\frac{p}{1-p}\right]$$

and α_k, β_k and λ_k are the model parameters.

In (32) the logit is assumed to be a polynomial of degree at most a in the covariates score X and this polynomial is possibly different for each of the four combinations of S and G. Simplification of this general model is achieved by data analysis in which various submodels of (32) are examined. Polynomials in X of degree 2 or more may be used to allow for curvilinear logit functions. For example, in the PALLID/ASHEN example the following logistic regression models were found to give satisfactory fits to the data in which X is the operational SAT-V score that does not include the studied item.

ASHEN, White examinees:

$$\text{logit } P(Y_S = 1 | S = c, G = r, \text{SAT-V})$$
$$= -1.970 - 0.458\ (\text{SAT-V}) + 0.581\ (\text{SAT-V})^2.$$

PALLID, White examinees:

$$\text{logit } P(Y_S = 1 | S = t, G = r, \text{SAT-V})$$
$$= -3.885 + 3.081\ (\text{SAT-V}) - 1.184\ (\text{SAT-V})^2 + 0.255\ (\text{SAT-V})^3.$$

ASHEN, Hispanic examinees:

$$\text{logit } P(Y_S = 1 \mid S = c, G = f, \text{SAT-V})$$

$$= -0.621 - 1.907\ (\text{SAT-V}) + 0.917\ (\text{SAT-V})^2.$$

PALLID, Hispanic examinees:

$$\text{logit } P(Y_S = 1 \mid S = t, G = f, \text{SAT-V})$$

$$= -1.194 + 0.174\ (\text{SAT-V}) + 0.273\ (\text{SAT-V})^2.$$

Let $\tilde{p}(j, g, x)$ denote the estimated conditional probability (or fitted probability) that results from the logistic regression analysis. The fitted probabilities are related to the estimated logits in (32) according to the following formula.

$$\tilde{L}(j, g, x) = \text{estimated logit } [(P(Y_S = 1 \mid S = j, G = g, X = x))]$$

then

$$\tilde{p}(j, g, x) = \exp(\tilde{L}(j, g, x))/(1 + \exp(\tilde{L}(j, g, x))).$$

The four fitted probability functions for the estimated logits given above are displayed in Figure 2. We see that the predicted probabilities for the PALLID item for the Hispanic group are quite different from the other three.

Insert Figure 2 about here

Once a satisfactory logistic regression model is selected, we may use it to obtain various covariate adjusted estimates. The fitted probabilities, $\tilde{p}(j, g, x)$ are estimates of the conditional probability,

$$p(j, g, x) = P(Y_S = 1 \mid S = j, G = g, X = x). \tag{33}$$

Define $\tilde{p}_j$ by

$$\tilde{p}_j = \sum_{x,g} \tilde{p}(j, g, x)\, n_{jgx} / \left(\sum_{x,g} n_{jgx} \right), \tag{34}$$

where n_{jgx} is the number of examinees in the study with $S = j$, $G = g$ and $X = x$.

Thus, $\tilde{p}_j$ may be viewed as an estimate of p_j, the proportion of examinees in the population who give the predicted response to item j in the pair (t, c), that is based on the smoothed predicted probabilities, $\tilde{p}(j, g, x)$. However, if the submodel of (32) that is selected to represent the data contains α_0 and β_0 as free parameters it may be shown that $\tilde{p}_j$ and the raw proportion, $\hat{p}_j$ are equal. Because we allowed α_0 and β_0 to be free in our analysis, covariance adjustment does not change our estimate of the main effect parameter.

Estimating T

The interaction parameter T defined in (8) can be estimated directly or by the use of covariate adjustments. Let $\tilde{p}_{jg}$ denote the proportion of examinees making the predicted response among all those exposed to item $j = t$ or c in group g $(g = f, r)$. The argument given in the earlier section shows that the difference of sample differences in proportions,

$$\hat{T} = \hat{p}_{tf} - \hat{p}_{cf} - (\hat{p}_{tr} - \hat{p}_{cr}), \tag{35}$$

estimates the ACE-differences, T. In the PALLID/ASHEN example the four proportions that make up $\hat{T}$ are given below.

	(*t*) PALLID	(*c*) ASHEN
Whites (*r*)	.51	.50
Hispanics (*f*)	.56	.36

the value of $\hat{T}$ is therefore

$$\hat{T} = (.56 - .36) - (.51 - .50) \tag{36}$$

$$= .19.$$

A covariate adjusted estimate of T can also be obtained from the fitted probabilities resulting from a logistic regression analysis. Let $\tilde{p}_{jg}$ be defined by

$$\tilde{p}_{jg} = \sum_x \tilde{p}(j, g, x)\, n_{jgx} / (\sum_x n_{jgx}) \tag{37}$$

where $\tilde{p}(j, g, x)$ and n_{jgx} are as defined earlier. Then the covariate adjusted estimate of T is

$$\tilde{T} = \tilde{p}_{tf} - \tilde{p}_{cf} - (\tilde{p}_{tr} - \tilde{p}_{cr}). \tag{38}$$

If the submodel of (32) that is selected to represent the data contains α_0, β_0, γ_0 and λ_0 as free parameters then it may be shown that $\tilde{p}_{jg}$ and $\hat{p}_{jg}$ are equal. Because we have done this in the models fit to the data in the PALLID/ASHEN example, our estimates, $\tilde{T}$ and $\hat{T}$, are equal.

Estimating $T(x)$

When sample sizes are very large, a useful direct estimate of $T(x)$ is available. In analogy with (35) it is

$$\hat{T}(x) = \hat{p}_{tfx} - \hat{p}_{cfx} - (\hat{p}_{trx} - \hat{p}_{crx}) \tag{39}$$

where $\hat{p}_{jgx}$ is the proportion of examinees who made the predicted response among all those for whom $S = j$, $G = g$ and $X = x$. However, in practice, where samples are often small, (39) yields very noisy estimates of $T(x)$ that can mask trends. Instead, a more useful approach is to use the fitted probabilities from the logistic regression analysis, $\tilde{p}(j, g, x)$. This yields

$$\tilde{T}(x) = \tilde{p}(t, f, x) - \tilde{p}(c, f, x) - (\tilde{p}(t, r, x) - \tilde{p}(c, r, x)). \tag{40}$$

When x is a univariate score, a graph of $\tilde{T}(x)$ versus x is a useful summary of the results for items t and c of the randomized DIF study. Figure 3 shows a plot of $T(x)$ versus x for the PALLID/ASHEN example in which the covariate is the SAT-V score.

Insert Figure 3 about here

Estimates of T_w are easily derived from (40) via the formula

$$\tilde{T}_w = \sum_x \tilde{T}(x)\, w(x) \tag{41}$$

for any set of weights $w(x)$. In particular, when

$$w(x) = \frac{\sum_j n_{jfx}}{\sum_{j,x} n_{jfx}} \tag{42}$$

We obtain an estimate of T_f, $T(x)$ weighted by the distribution of X in the focal group. It is

$$\tilde{T}_f = \sum_x \tilde{T}(x) \frac{\sum_j n_{jfx}}{\sum_{j,x} n_{jfx}} . \tag{43}$$

For the PALLID/ASHEN example $\tilde{T}_f$ is .17, which agrees with the difference in standardization parameters reported in Schmitt et al. (1988). This agreement is due to several factors. Most importantly the sample size for the Hispanic groups who responded to the t and c items were sufficiently large (1,619 and 1,522, respectively) that the distribution of the covariate scores in these two groups were similar to the distribution obtained by pooling them. In addition, the curves reported in Figure 2 are the result of careful data analysis and represent the noisy raw proportions in the data very well. Finally, the estimate of the standardization parameter is based on an external matching criterion that is the same as that used in the logistic regression analyses reported here--the SAT-V score. We note that the use of an external matching criterion that does not include the studied item is generally not an appropriate procedure for measuring the amount of DIF exhibited in an item, but in this case it is appropriate since the parameter of

ultimate interest is the average interaction parameter, T_j, given in (18), rather than the DIF values above.

SUMMARY OF STEPS FOR USING LOGISTIC REGRESSION IN A RANDOMIZED DIF STUDY

The theory and practice of logistic regression are now fairly well established. The discussions in Cox (1970) and in Hosmer and Lemeshow (1989) are very helpful and software is available in the SAS, SPSS and BMDP packages. We suggest the following checklist for the use of logistic regression in the analysis of data from randomized DIF studies:

- Be sure that the variables used as covariate scores are, in fact, covariates--i.e., they are unaffected by whether the examinee was exposed to the *t* or *c* item.

- Consider including as many covariate scores as possible in the analysis-- e.g., math as well as verbal scores, or subscores such as rights and omits on formula-scored tests.

- Consider including powers higher than linear or quadratic terms in order to improve the fit of the model.

- Start with a large model like that in (32), and then simplify it to the point where there are as few parameters as possible without a degradation in the fit.

- Check the fit of the model in at least the following two ways:

  a) See if the inclusion of a term in the model adds substantially to its fit as measured by the standard one-degree-of-freedom likelihood ratio test

  b) Plot the fitted proportions from the model along with the observed proportions as functions of the covariate scores for each combination of group (f or r) and item (t or c). The fitted proportions should go through the middle of the scatter of observed proportions.

- In addition, check residuals from the model for outliers, remove them to see if they are responsible for unusual features of the resulting model.
- Remember that the point of this careful data analysis is to find a smooth function of the covariate score(s) that adequately smoothes the noisy observed proportions, $\hat{p}_{jgx}$.

- Use the fitted proportions, $\tilde{p}(j, g, x)$, to compute $\tilde{T}(x)$ and single number summaries like $\tilde{T}_f$.

- Do not concentrate on interpreting the coefficients of the finally selected logistic regression model, i.e. $\hat{\alpha}_k$, $\hat{\beta}_k$, $\hat{\gamma}_k$ or $\hat{\lambda}_k$, because these are in the logit scale. Rather,

compare the four functions $\tilde{p}(t, r, x)$, $\tilde{p}(c, r, x)$, $\tilde{p}(t, f, x)$ and $\tilde{p}(c, f, x)$ via plots, as in Figure 2, and interpret these differences.

5.3 DIFFERENTIAL SPEEDEDNESS ASSESSED UNDER CONTROLLED CONDITIONS

An example of a special confirmatory study where the location of the items was varied is the Dorans, Schmitt and Curley (1988) study. This study examined directly how differential speededness affects the magnitude of DIF. In addition, it evaluated how well the procedure of excluding not reached examinees from the calculation of the DIF statistic adjusts for the effects of differential speededness. The purpose of the study was to answer two questions:

- Does an item's DIF value depend upon its location in the test?
- If so, can the item location effect be removed via a statistical adjustment of the DIF statistic?

For a detailed description of the study see Dorans, Schmitt and Curley (1988).

For the purposes of their study, one non-operational 45-item and one non-operational 40-item SAT-Verbal pretest were labelled "Form A" and "Form B", respectively. The ten analogy items appearing in Form A in positions 36 to 45 were combined with the antonyms, sentence completions, and reading comprehension items from Form B to create "Form C", a 40-item section in which the ten analogies appeared in positions 16 to 25. Similarly, the analogy items from Form B in positions 16 to 25 were combined with the antonyms, sentence completions, and reading comprehension items from Form A to create "Form D", a 45-item test in which the ten analogies were shifted to the end of the section in positions 36 to 45.

This particular design afforded an opportunity to examine how differential speededness for Blacks affects the magnitude of DIF statistics on two sets of analogy items. As a control analysis, Dorans, Schmitt and Curley (1988) also conducted differential speededness and DIF analyses for females on the same sets of analogy items. Standardized distractor analyses (Dorans & Holland, in press) that focused on not reached were used to assess differential speededness.

Figures 4 and 5 depict the degree of differential speededness observed for Blacks and females on the Form A and Form D analogy items, respectively. In these figures, the STD P-DIF(NR) values, in percentage units, are plotted against item number. Absolute values of 5% or greater indicate a sizeable degree of differential speededness. A positive STD P-DIF(NR) value means that the focal group, Blacks or females, is not reaching the item to the degree that the base or reference group, Whites or males, is. Conversely, a negative STD P-DIF(NR) value means the focal group is reaching the item in greater proportions than the matched base group.

Insert Figures 4 and 5 about here

In Figure 4, there is little evidence of differential speededness for females. For Blacks, there is some evidence, particularly on items 42 and 43, and possibly 40 and 41. In Figure 5, for females, item 44 is approaching the 5% cutoff. For Blacks, differential speededness is quite pronounced. Items 41, 42, 43, and 44 are at or above the 5% value, while items 38, 39, and 40 are approaching the 5% value. Note that across Figures 4 and 5 all but one item has a positive STD P-DIF(NR) value for Blacks, indicating that Blacks reach items at the end of the 45-item Verbal 1 section at a slower rate than a matched group of Whites, as reported by

Dorans, Schmitt and Bleistein (1988). In contrast, the STD P-DIF(NR) values for females are either at 0 (9 of 20 items) or slightly negative, indicating that females get further into the test than a matched group of males.

There are no figures for the analogy items on Forms B or C because there is no differential speededness on the analogy items in positions 16 to 25 of the 40-item format. In fact, all examinees reached these items.

A major goal of the Dorans, Schmitt and Curley research was to ascertain whether or not there was a position effect on DIF statistics. Evidence has been presented for a differential speededness effect for Blacks, and of negative DIF, predominantly for Blacks on the earlier, easier analogy items. In addition item position effects were reported.

Does an item's DIF value depend on its location in the test? Dorans, Schmitt and Curley (1988) reported that the answer is yes for some items, particularly when one position is subject to a differential speededness effect while the other is not.

The second question to be addressed by the Dorans, Schmitt and Curley research was: Can the item location effect be removed via a statistical adjustment? In particular, does exclusion of the candidates who do not reach the item from calculation of the DIF statistic produce a statistic that is less sensitive to position? All things considered, the adjustment for not reached tended to dampen the position effect for most items. It did not, however, statistically remove completely the speededness effect.

6. SUMMARY

This paper provided prescriptions for the practice of conducting research into the evaluation of DIF hypotheses. Advice was given for both observational studies with operational item data and controlled studies with specially constructed items. The following checklist can be used to guide the conduct of observational studies.

SUMMARY OF STEPS IN THE EVALUATION OF DIF HYPOTHESES USING OBSERVATIONAL DATA

- Operationalize the definition of the postulated DIF factors in order to permit the objective classification of items.
- Classify all items in accordance with postulated DIF factors.
- Define the appropriate focal and reference groups.
- Select appropriate samples.
- Determine the matching criterion considering dimensionality, reliability, and criterion refinement issues.
- Determine what statistical adjustments are relevant (e.g., speededness and omission).
- Select an appropriate DIF estimate based on the above considerations.
- Calculate DIF statistics for the key, distractors, and response style factors.
- Evaluate relevant information provided by distractor and difference plots.
- Summarize DIF information by the postulated factors; use descriptive statistics (e.g., correlate comparable DIF outcomes with hypothesized factors using appropriate statistical methods).

- Determine whether the DIF information supports the hypothesized DIF factors.

Section 4.2 described the rudiments of a theory of causal inference, the success of which hinges on putting the J. S. Mill quote from nearly 150 years ago into action by measuring causation through experimental manipulation. Section 5.2 describes the specifics of one approach towards a accomplishing this. The following checklist can be used to guide future randomized DIF studies.

SUMMARY OF STEPS IN THE CONFIRMATORY EVALUATION OF DIF HYPOTHESES USING SPECIALLY CONSTRUCTED ITEMS

- Construct sets of items (treatment and control) in accordance with postulated DIF factors; control extraneous factors to the extent possible.
- Define the focal and reference groups and randomly determine control and treatment subgroups.
- Select appropriate sample sizes; replicate administrations when needed in order to obtain sufficient sample sizes.
- Determine the matching criterion that is a covariate in the sense used here; use an external matching criterion when possible. Consider dimensionality and criterion refinement issues.
- Specify what statistical adjustments are relevant (e.g., speededness and omission).
- Calculate DIF statistics for the key, distractors, and response style factors.
- Evaluate relevant information provided by distractor and difference plots.
- Summarize DIF information by the postulated factors; use descriptive and inferential statistics.

- Determine whether the DIF information supports the hypothesized DIF factors.

These randomized DIF studies are distinguished by the careful construction of hypothesis items and their controls, the control of extraneous factors, the use of randomization, and the quest for adequate samples to achieve enough statistical power to detect affects related to the DIF hypotheses. If these conditions are met in practice, then DIF findings, if replicated, may suggest changes in educational assessment and practice. Evaluation of DIF hypotheses is complicated however by a variety of practical and ethical considerations. Sound scientific method needs to operate within these constraints and achieve success in advancing knowledge that will affect test development practice, assessment, and educational practice.

REFERENCES

Bleistein, C. A., Schmitt, A. P., & Curley, W. E. (1990, April) Factors hypothesized to affect the performance of Black examinees on SAT-Verbal analogy items. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Bleistein, C. A., & Wright, D. (1987). Assessment of unexpected differential item difficulty for Asian-American candidates on the Scholastic Aptitude Test. In A. P. Schmitt & N. J. Dorans (Eds.), Differential item functioning on the Scholastic Aptitude Test (RM-87-1). Princeton, NJ: Educational Testing Service.

Cox, D. R. (1970). Analysis of binary data. London: Methuen.

Donoghue, J. R., Holland, P. W., and Thayer, D. T. (in press). A monte-carlo study of factors that affect the Mantel-Haenszel and standardization measure of differential item functioning. In P. W. Holland and H. Wainer (Eds.) Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J. (1982). Technical review of item fairness studies: 1975-1979 (SR-82-90). Princeton, NJ: Educational Testing Service.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. Applied Measurement in Education, 3, 217-233.

Dorans, N. J., & Holland, P. W. (in press). DIF detection and description: Mantel-Haenszel and Standardization. In P. W. Holland and H. Wainer (Eds.) Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Kulick, E. (1983). Assessing unexpected differential item performance of female candidates on SAT and TSWE Forms administered in December 1977: An application of the standardization approach (RR-83-9). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing differential item functioning on the Scholastic Aptitude Test. Journal of Educational Measurement, 23, 355-368.

Dorans, N. J., & Lawrence, I. M. (1987). The internal construct validity of the SAT (RR-87-35). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1988). The standardization approach to assessing differential speededness (RR-88-31). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Schmitt, A. P., & Curley, W. E. (1988, April). Differential speededness: Some items have DIF because of where they are, not what they are. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Echternacht, G. (1972). An examination of test bias and response characteristics of six candidate groups taking the ATGSB (RR-72-4). Princeton, NJ: Educational Testing Service.

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. Journal of Educational Measurement, 26, 147-160.

Holland, P. W. (1985). On the study of differential item performance without IRT. Proceedings of the 27th Annual Conference of the Military Testing Association, San Diego, CA, Vol. I, pp. 282-287.

Holland, P. W. (1986). Statistics and Causal Inference. Journal of the American Statistical Association, 81, 945-970.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test Validity. Hillsdale, NJ: Erlbaum.

Hosmer, D. W. & Lemeshow, S. (1989). Applied logistic regression. New York: Wiley.

Hunter, R. V., & Slaughter, C. D. (1980). ETS test sensitivity review process. Princeton, NJ: Educational Testing Service.

Kulick, E. (1984). Assessing unexpected differential item performance of Black candidates on SAT form CSA6 and TSWE form E33 (SR-84-80). Princeton, NJ: Educational Testing Service.

Mill, J. S. (1843). A system of logic.

Ramsey, P. (in press). Sensitivity review process. In P. W. Holland and H. Wainer (Eds.) Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rogers, J., Dorans, N. J., & Schmitt, A. P. (1986). Assessing unexpected differential item performance of Black candidates on SAT form 3GSA08 and TSWE form E43 (SR-86-22). Princeton, NJ: Educational Testing Service.

Rogers, J., & Kulick, E. (1987). An investigation of unexpected differences in item performance between Blacks and Whites taking the SAT. In A. P. Schmitt & N. J. Dorans (Eds.), Differential item functioning on the Scholastic Aptitude Test (RM-87-1). Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66, 688-701.

Scheuneman, J. D. (1978). Ethnic group bias in intelligence test items. In S. W. Lundsteen (Ed.), Cultural factors in learning and instruction. New York: ERIC Clearinghouse on Urban Education, Diversity Series, No. 56.

Scheuneman, J. D. (1981). A response to Baker's criticism. Journal of Educational Measurement, 16, 143-152.

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. Applied Measurement in Education, 2, 255-275.

Schmitt, A. P. (1985). Assessing unexpected differential item performance of Hispanic candidates on SAT form 3FSA08 and TSWE form E47 (SR-85-169). Princeton, NJ: Educational Testing Service.

Schmitt, A. P. (1988). Language and cultural characteristics that explain differential item functioning for Hispanic examinees on the Scholastic Aptitude Test. Journal of Educational Measurement, 25, 1-13.

Schmitt, A. P., & Bleistein, C. A. (1987). Factors affecting differential item functioning for Black examinees on Scholastic Aptitude Test analogy Items (RR-87-23). Princeton, NJ: Educational Testing Service.

Schmitt, A. P., Curley, W. E., Bleistein, C. A., & Dorans, N. J. (1988, April). Experimental evaluation of language and interest factors related to differential item functioning for Hispanic examinees on the SAT-Verbal. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. Journal of Educational Measurement, 27, 67-81.

Schmitt, A. P., & Dorans, N. J. (in press). Factors related to differential item functioning for Hispanic examinees on the Scholastic Aptitude Test. In J. Deneeu, G. Keller & R. Magallan (Eds.), Assessment and Access: Hispanics in Higher Education. New York: Sunny Press.

Schmitt, A. P., Dorans, N. J., Crone, C. R. & Maneckshana, B. T. (1991). Differential speededness and item omit patterns on the SAT (RR-91-50). Princeton, NJ: Educational Testing Service.

Shepard, L. A. (1987). Discussant comments on the NCME Symposium: Unexpected differential item performance and its assessment among Black, Asian-American, and Hispanic students. In A. P. Schmitt & N. J. Dorans (Eds.), Differential item functioning on the Scholastic Aptitude Test (RR-87-1). Princeton, NJ: Educational Testing Service.

Thissen, D., Steinberg, L., & Wainer, H. (in press). Detection of differential item functioning using the parameters of item response models. In H. Wainer and P. W. Holland (Eds.), Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates.

Stricker, L. J. (1982). Identifying test items that perform differently in population subgroups: A partial correlation index. Applied Psychological Measurement, 6, 261-273.

Wright, D. (1987). An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance. In A. P. Schmitt & N. J. Dorans (Eds.), Differential item functioning on the Scholastic Aptitude Test (RM-87-1), Princeton, NJ: Educational Testing Service.

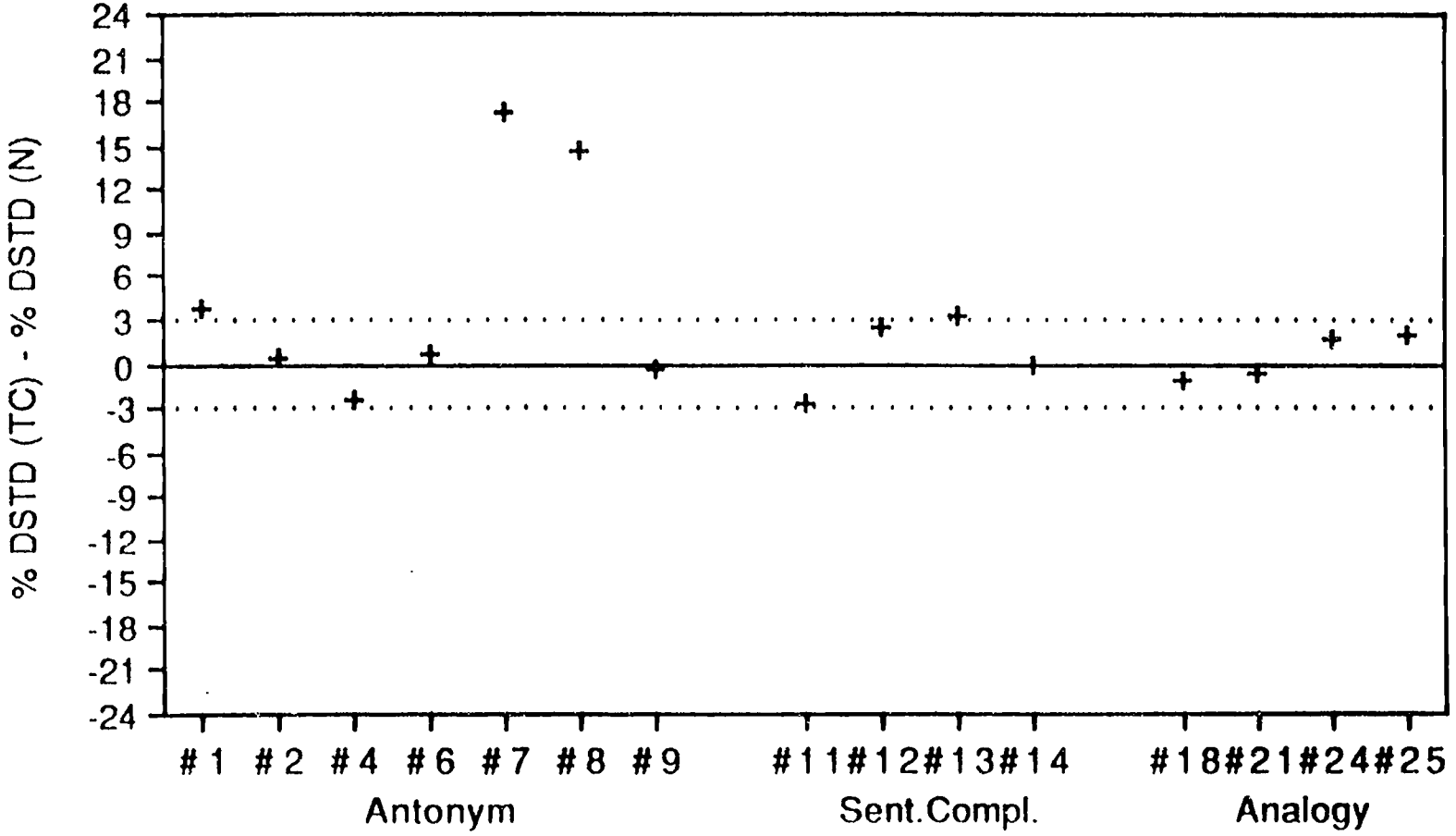

Figure 1 : True Cognate Item Effects for All Hispanics

Figure 2:

FITTED PROBABILITIES FOR THE "PALLID/ASHEN" ITEM PAIR REFERENCE GROUP=WHITE, FOCAL GROUP=HISPANIC

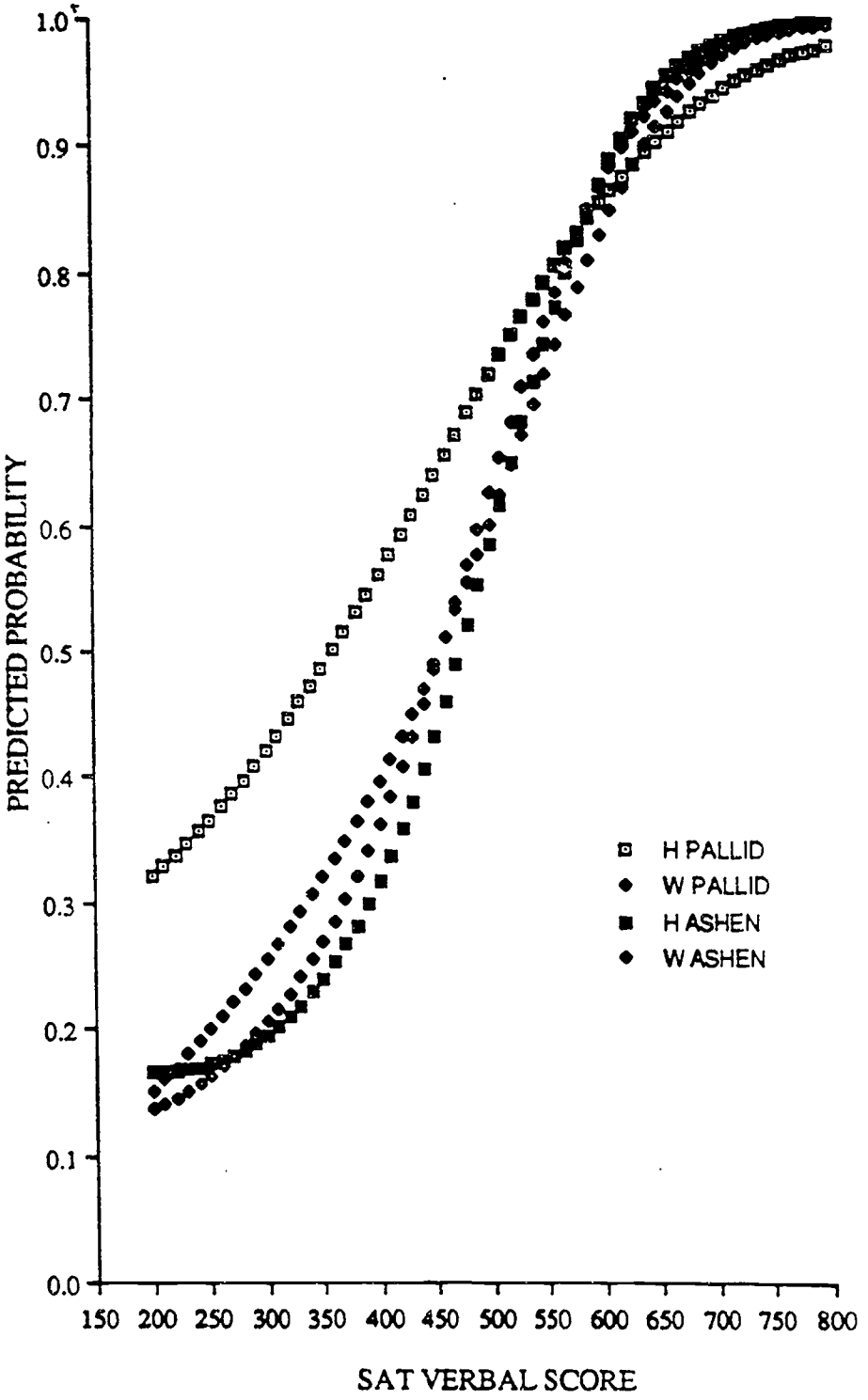

Figure 3:
PLOT OF T(X) VS X
FOR X=SAT VERBAL SCORE
FOR "PALLID/ASHEN" ITEM PAIR

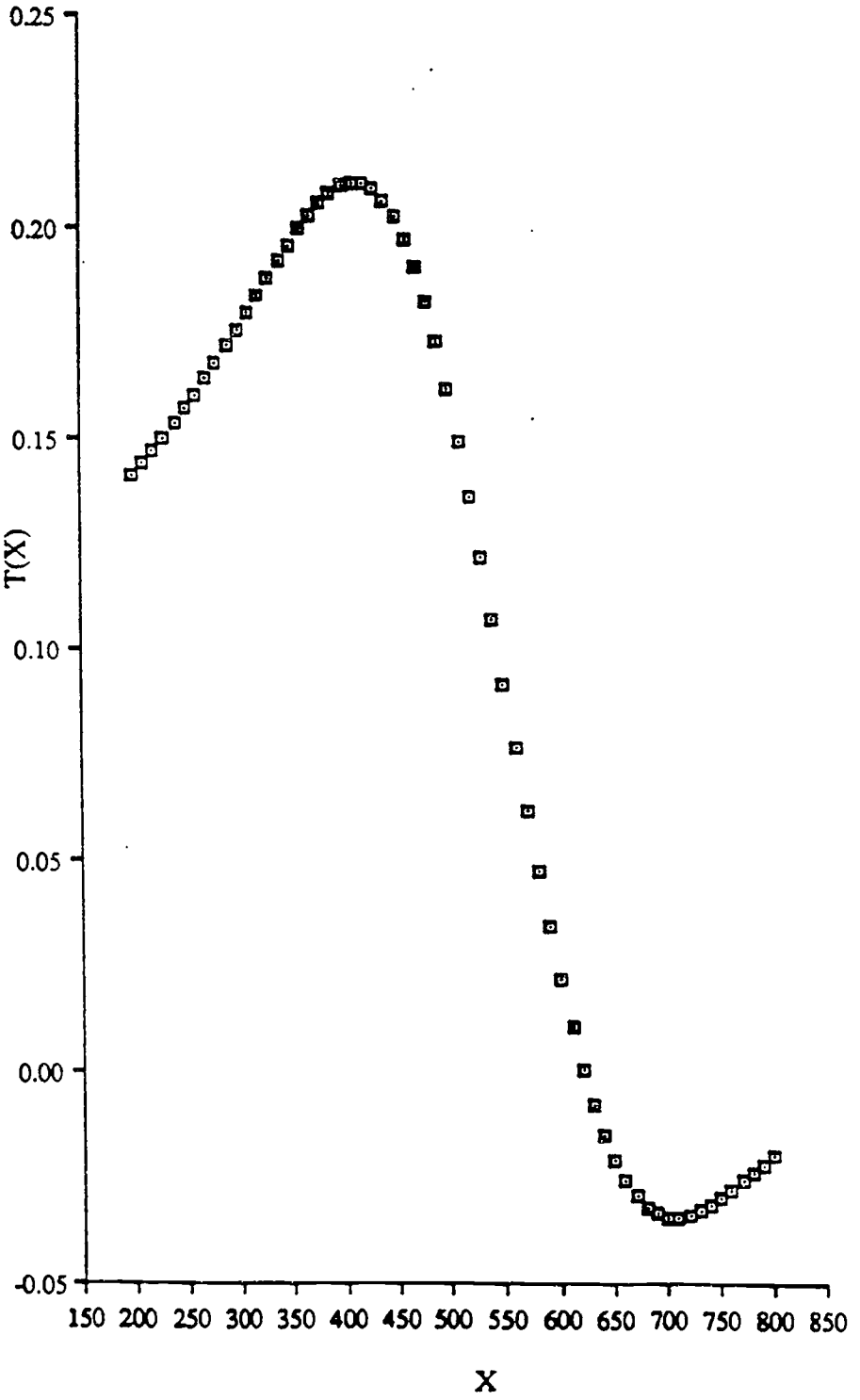

Figure 4: Differential Speededness On Form A

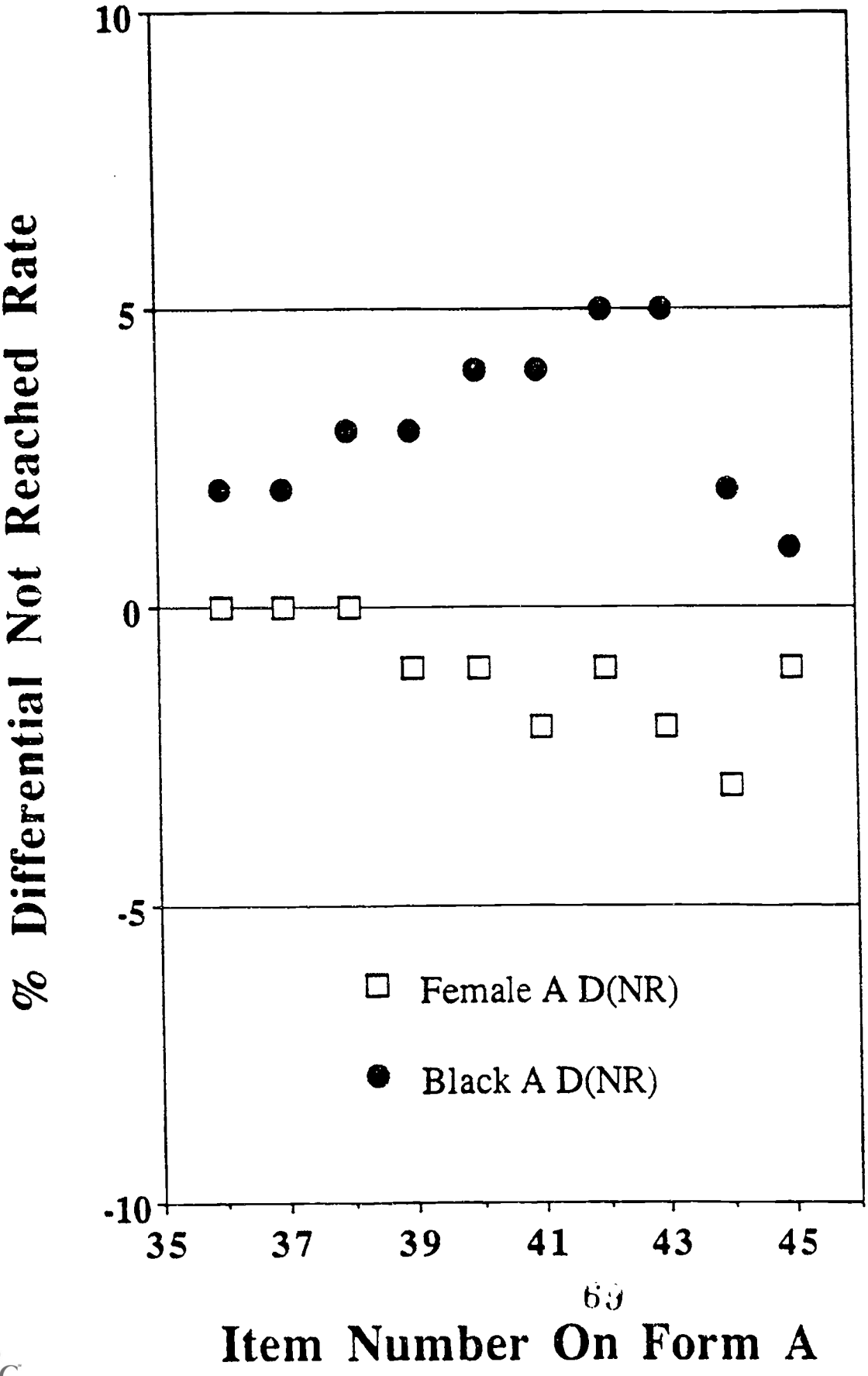

Figure 5: Differential Speededness On Form D